100 Years of Digital Data

Georgia Institute of Technology

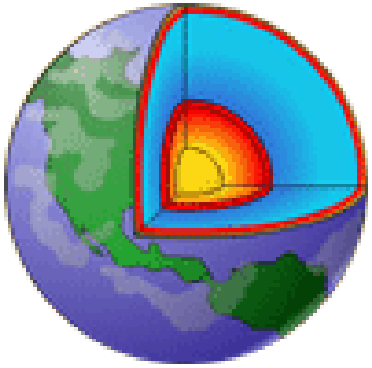
Dr. Francine Berman

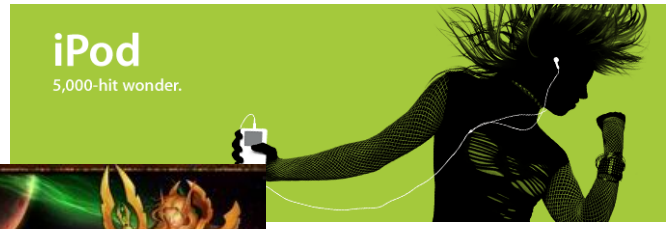
Director, San Diego Supercomputer Center

Professor and High Performance Computing Endowed Chair, UC San Diego

Digital Data Drives the Information Age

Education





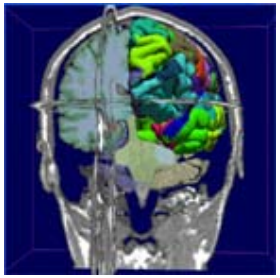
Entertainment

Shopping

Information

Google™

Health



Business

How much Digital Data is there?

- 5 exabytes of digital information produced in 2003
- **161 exabytes of digital information produced in 2006**
  - 25% of the 2006 digital universe is born digital (digital pictures, keystrokes, phone calls, etc.)
  - 75% is replicated (emails forwarded, backed up transaction records, movies in DVD format)
- **1 zettabyte aggregate digital information projected for 2010**

Kilo	10^3
Mega	10^6
Giga	10^9
Tera	10^{12}
Peta	10^{15}
Exa	10^{18}
Zetta	10^{21}

SDSC HPSS tape archive = **25+ PetaBytes**

iPod (up to 20K songs) = **80 GB**

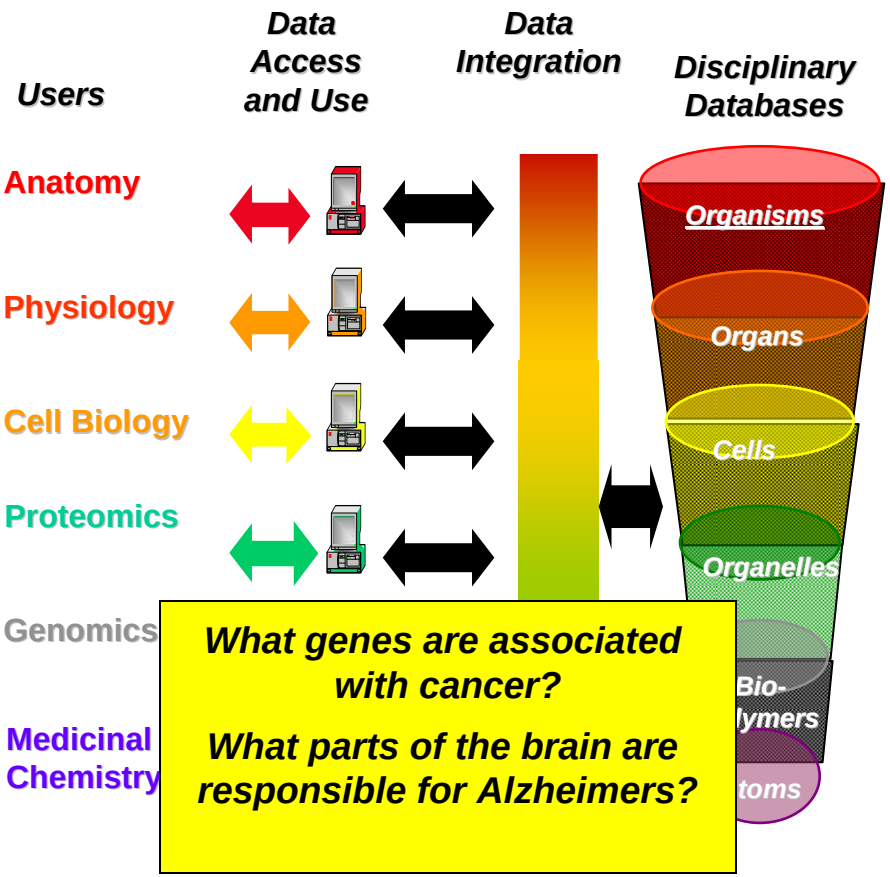
1 novel = **1 MegaByte**

U.S. Library of Congress manages **295 TB** of digital data, 230 TB of which is "born digital"

Source: *"The Expanding Digital Universe: A forecast of Worldwide Information Growth through 2010" IDC Whitepaper, March 2007*

Data Drives Research and Education

Data at multiple scales in the Biosciences



What genes are associated with cancer?

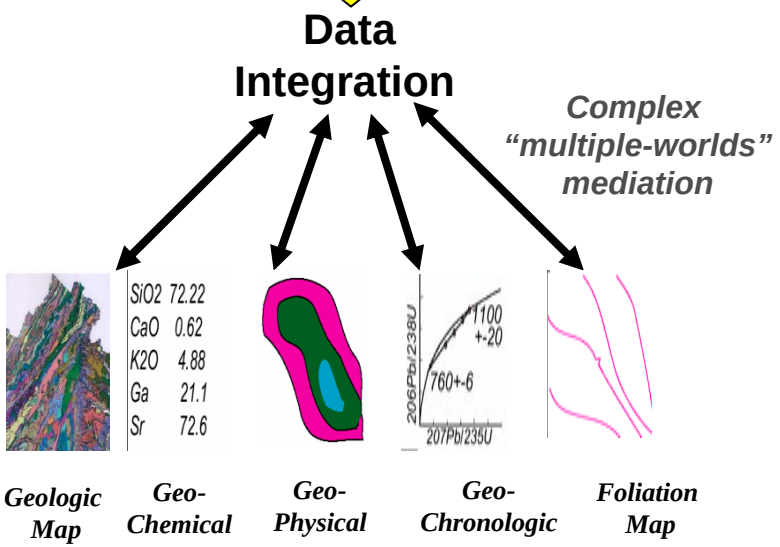
What parts of the brain are responsible for Alzheimers?

Data from multiple sources in the Geosciences

Where should we drill for oil?

What is the Impact of Global Warming?

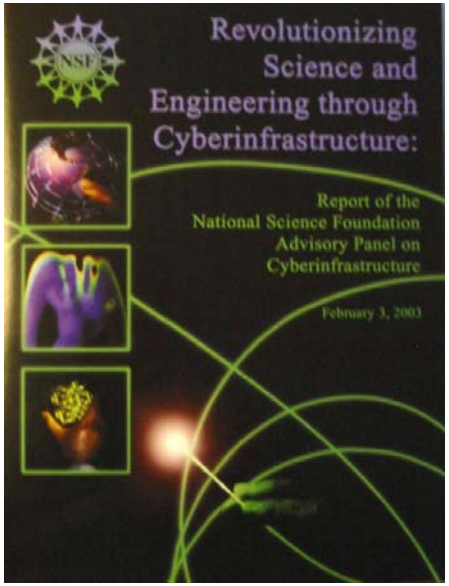
How are the continents shifting?



Images courtesy of Mark Miller and Chaitan Baru

Data a Fundamental Component of Cyberinfrastructure

- **Cyberinfrastructure** is the organized aggregate of technologies enabling access and coordination of information technology resources to facilitate science, engineering, and societal goals.
  - *Data*
  - *Computation*
  - *Communication*
  - *Visualization*
  - *Scientific Instruments*
  - *Expertise, etc.*



Published in 2003, NSF Blue Ribbon Panel (Atkins) Report provided a compelling and comprehensive vision of an integrated Cyberinfrastructure

Today's Presentation

- Revolutionizing astronomy through digital data
- Data Cyberinfrastructure Today – Designing and developing infrastructure to enable today's data-oriented applications
- Challenges in Building and Delivering Data Cyberinfrastructure

Astronomy in the Last Century

1908 **First reported detection of a magnetic field in any astronomical object** -- a sunspot. Discovered by George Hale using telescope at Mount Wilson Solar Observatory.

1916 Einstein introduces **General Theory of Relativity**

1916 Hubble shows that **galaxies exist outside the Milky Way Galaxy**

1930 **Discovery of Pluto** by Clyde Tombaugh (following work begun by Percival Lowell). Pluto's existence hypothesized based on anomalies in the orbits of Neptune and Uranus.

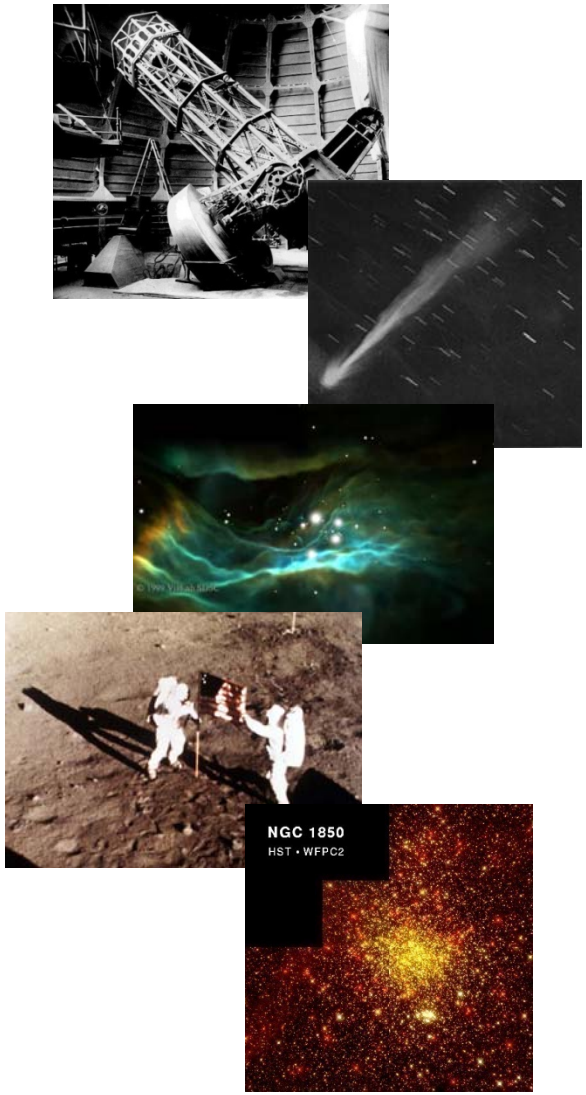
1957 Sputnik (first human-made satellite) launched, marking **beginning of the "Space Age"**

1958 **NASA created** (from former National Advisory Committee for Aeronautics) and other US Govt organizations

1965 Penzias and Wilson discover cosmic fossil radiation, providing **direct evidence of the Big Bang Theory**

1990 **Hubble Space Telescope** put into orbit

Etc., etc., etc.

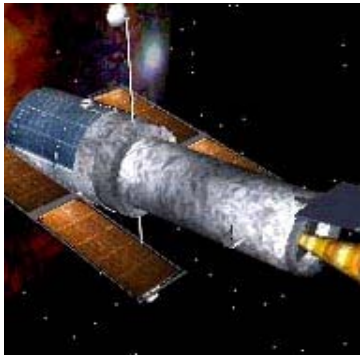


Information/images courtesy of UCAR, NVO, NASA

Astronomy Today

"The Universe is now being explored systematically, in a panchromatic way, over a range of spatial and temporal scales that lead to a more complete, and less biased understanding of its constituents, their evolution, their origins, and the physical processes governing them."

Towards a National Virtual Observatory



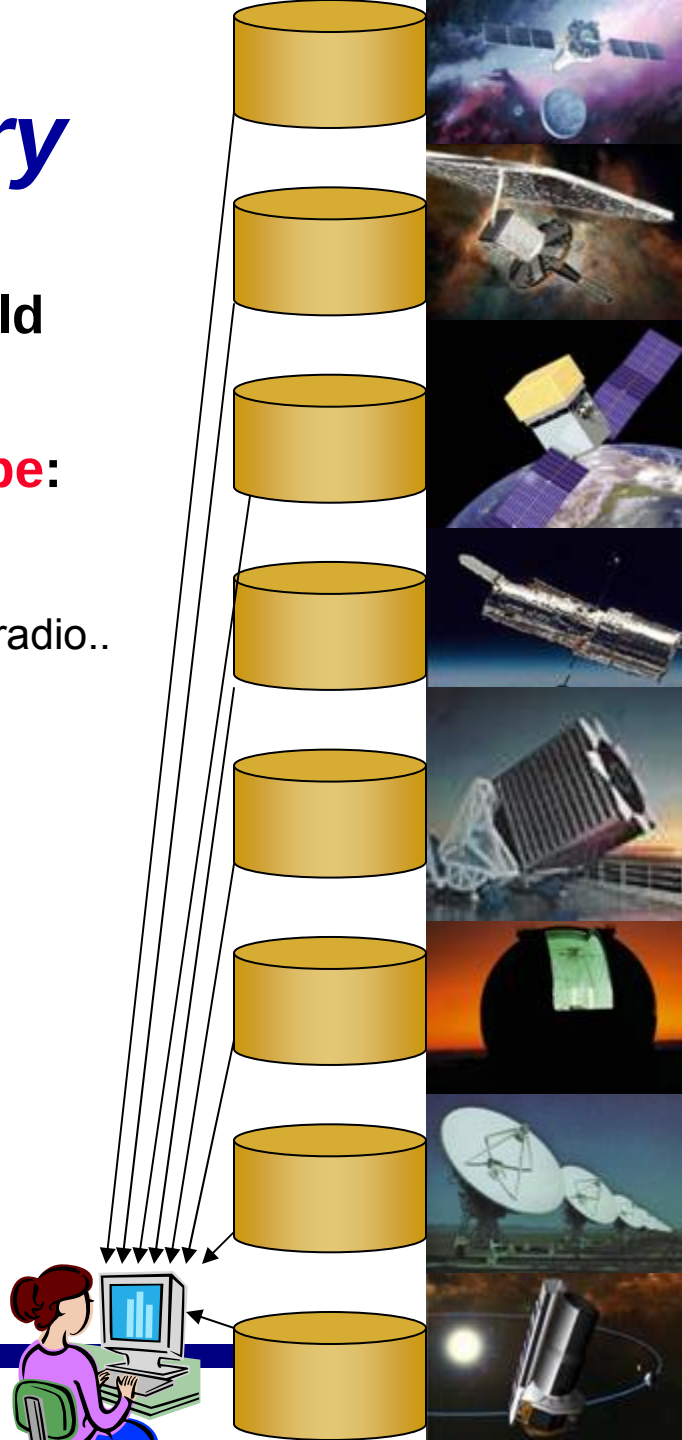
Hubble Telescope

Palomar Telescope

Sloan Telescope

The Virtual Observatory

- **Premise: most observatory data is (or could be) online**
- **So, the Internet is the world's best telescope:**
  - It has data on every part of the sky
  - In every measured spectral band: optical, x-ray, radio..
  - It's as deep as the best instruments
  - It is up when you are up
  - The "seeing" is always great
  - It's a smart telescope:
    links objects and data to literature on them
- **Software has became a major expense**
  - Share, standardize, reuse..

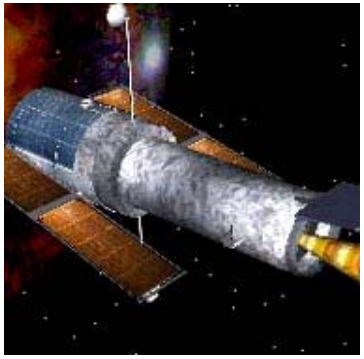


Downloading the Night Sky

The National Virtual Observatory

- NVO combines data from sky surveys and over 50 ground and space-based telescopes and instruments to create a comprehensive picture of the heavens

- HOW NVO Works
  - Raw data comes from large-scale synoptic telescopes. Scientists "clean" data, convert data from temporal to spatial, indexing over both dimmensions
  - NVO data available to the public without restriction after 1 year by community agreement
  - NVO databases distributed and mirrored at multiple sites



Hubble Telescope

Palomar Telescope

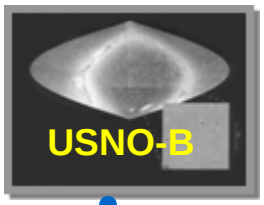
Sloan Telescope

NVO at SDSC

The Sloan Digital Sky Survey provides a 3D map of a million galaxies and quasars covering more than a quarter of the sky.

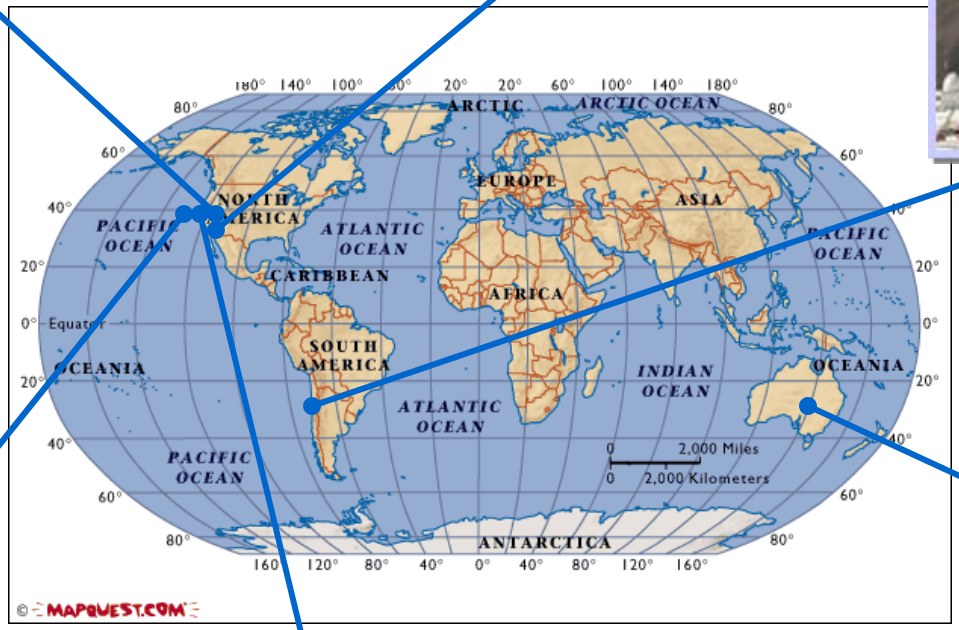
The 2 Micron All Sky Survey (2MASS) provides direct answers to questions on the large-scale structure of the Milky Way and the Local Universe

The Palomar Oschin telescope provides a catalogue of the entire northern sky in blue, red and near-infrared colors.



The USNO-B all-sky catalogue was obtained from various sky surveys during the last 50 years. USNO provides all-sky coverage and 85% accuracy for distinguishing stars from non-stellar objects

2MASS gathers data from a northern facility in Arizona and a southern facility in Chile

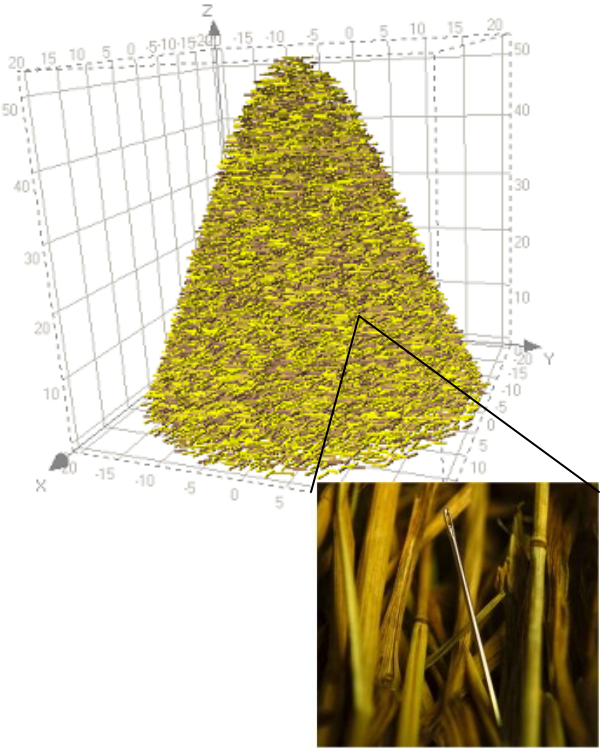


Photometric data from Mt. Stromlo observatory in Australia on several million stars gathered since 1992 to explore constitution of dark matter in the halo of the Milky Way

SDSC's NVO collection is nearly 100 TB and has grown over 5-fold since 2002

Data-Driven Astronomy

- **Looking for**
  - ***Needles in haystacks*** – the Higgs particle
  - ***Haystacks --*** Dark matter, Dark energy

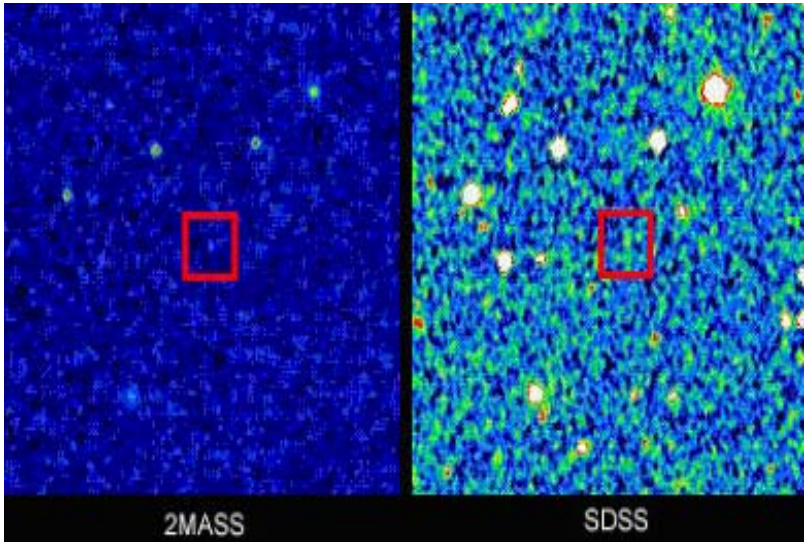


- **Statistical analysis often deals with**
  - Creating uniform samples
  - Data filtering
  - Assembling relevant subsets
  - Censoring bad data
  - “Likelihood” calculations
  - Hypothesis testing, etc.
- **Traditionally these are performed on files, most of these tasks are much better done inside a database**

Making Discoveries Using the NVO

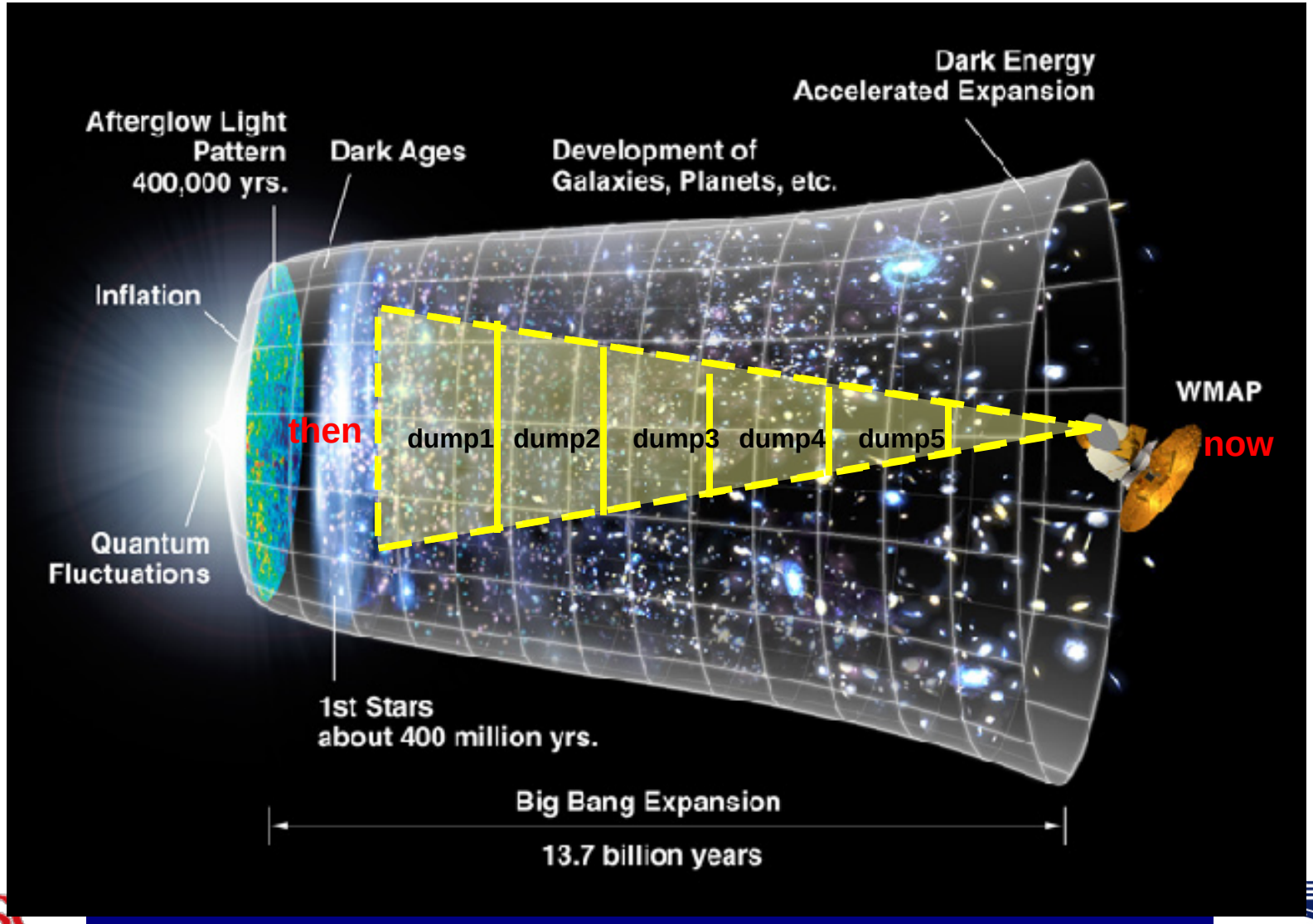
Scientists at Johns Hopkins, Caltech and other institutions confirmed the discovery of a new brown dwarf. Search time on 5,000,000 files went from months to minutes using NVO database tools and technologies.

Brown dwarfs are often called the "missing link" in the study of star formations. They are considered small, cool "failed stars".



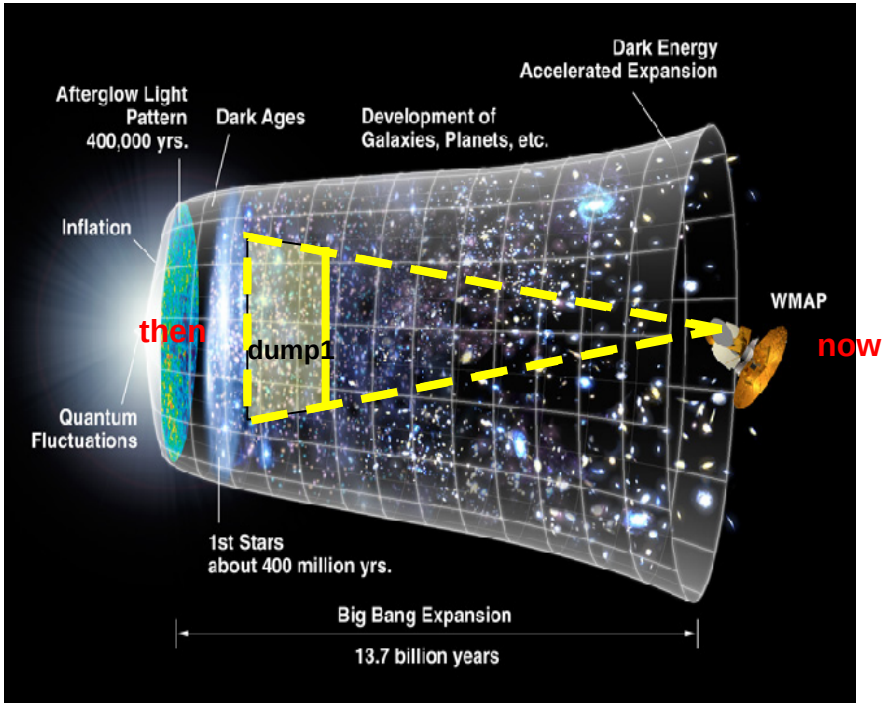
Evolving the Universe from the "Big Bang"

Composing simulation outputs from different timeframes builds up light-cone volume



After the "Big Bang" – the Universe's First Billion Years

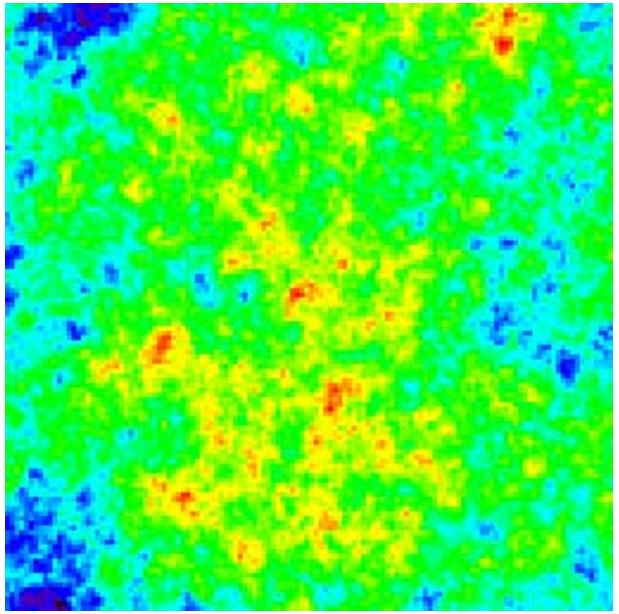
- **ENZO simulates the first billion years of cosmic evolution after the "Big Bang"**
- **Key period which represents**
  - A tumultuous period of intense star formation *throughout the universe*
  - Synthesis of the first heavy elements in massive stars
  - Supernovae, gamma-ray bursts, seed black holes, and the corresponding growth of supermassive black holes and the birth of quasars
  - Assembly of first galaxies



ENZO Simulations

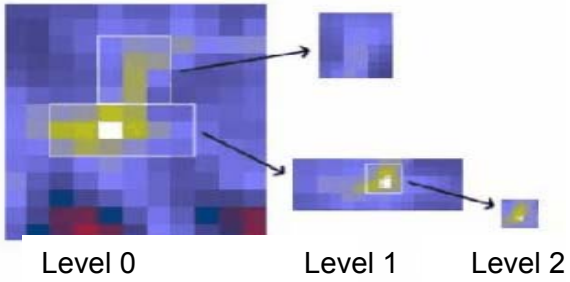
What ENZO does:

- Calculates the growth of cosmic structure from seed perturbations to form stars, galaxies, and galaxy clusters, including simulation of
  - *Dark matter*
  - *Ordinary matter (atoms)*
  - *Self-gravity*
  - *Cosmic expansion*
- Uses **adaptive mesh refinement** (AMR) to provide high spatial resolution in 3D
  - The Santa Fe light cone simulation generated over 350,000 grids at 7 levels of refinement
  - **Effective resolution = $65{,}536^3$**



Formation of a galaxy cluster

AMR



Computational Grand Challenge for Cosmology

ENZO at Petascale

- Self-consistent **radiation**-hydro simulations of structural, chemical, and radiative evolution of the universe simulates from first stars to first galaxies
- **Technical challenges:**
  - *Parallelizing the grid hierarchy metadata for millions of subgrids distributed across 10s of thousands of cores*
  - *Efficient dynamic load balancing of the numerical computations, taking memory hierarchy and latencies into account*
  - *Efficient parallel "packed AMR" I/O for 100 TB data dumps*
  - *Inline data analysis/viz. to reduce I/O*

Verifying Theory with Observation

- **James Webb Space Telescope** (JWST), coming in 2013 will probe the first billion years of the universe – providing observations of unprecedented depth and breadth
- Data will enable tight integration of observation and theory, and will enable simulations to approach realistic complexity
- **Petascale computing** *and* **scientific data management** essential for achieving new results

Slide modified from Mike Norman, image by Robert Harkness

Building and Delivering infrastructure for Data-oriented Applications

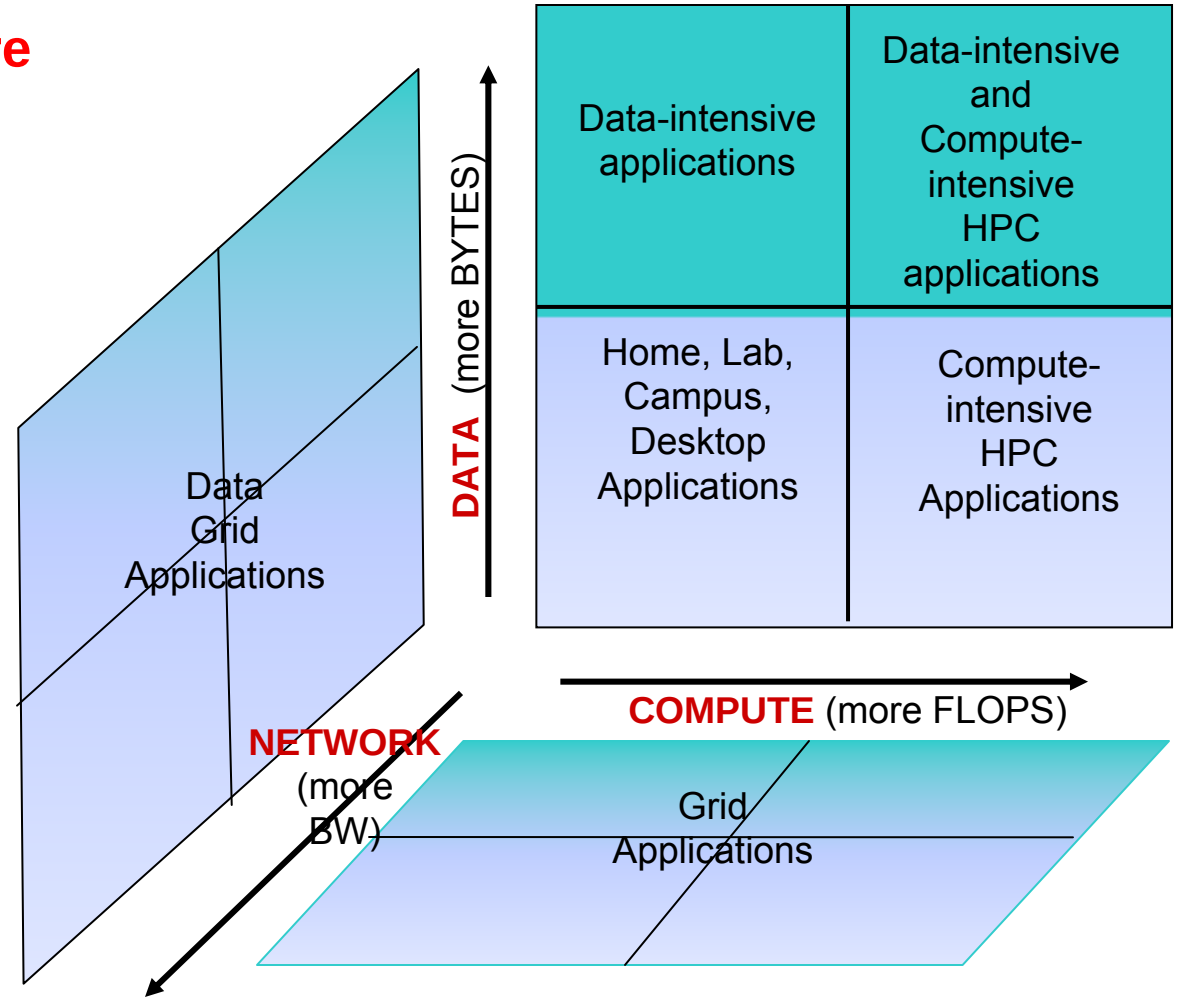
Today's Data-oriented Applications Span the Spectrum

Designing Infrastructure for Data:

Data and High Performance Computing

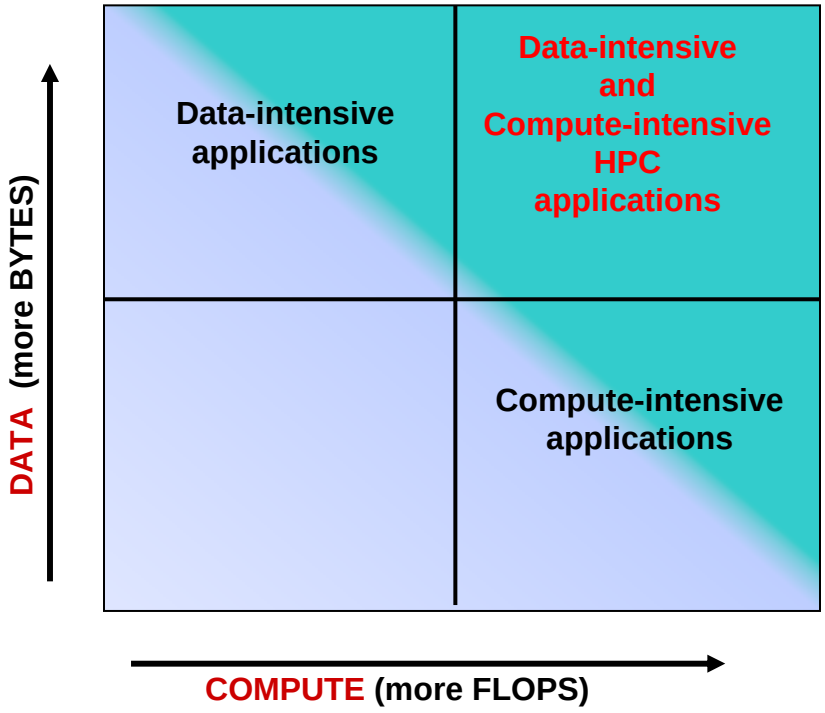
Data and Cyberinfrastructure Services

Support for management and preservation of data of community value



Data and High Performance Computing

- **For many applications, development of "balanced systems" needed to support applications which are both data-intensive and compute-intensive. Codes for which**
  - Grid platforms not a strong option
  - Data must be local to computation
  - I/O rates exceed WAN capabilities
  - Continuous and frequent I/O is latency intolerant
- **Scalability is key**
  - Need high-bandwidth and large-capacity local parallel file systems, archival storage



Data and HPC: What you see is what you've measured

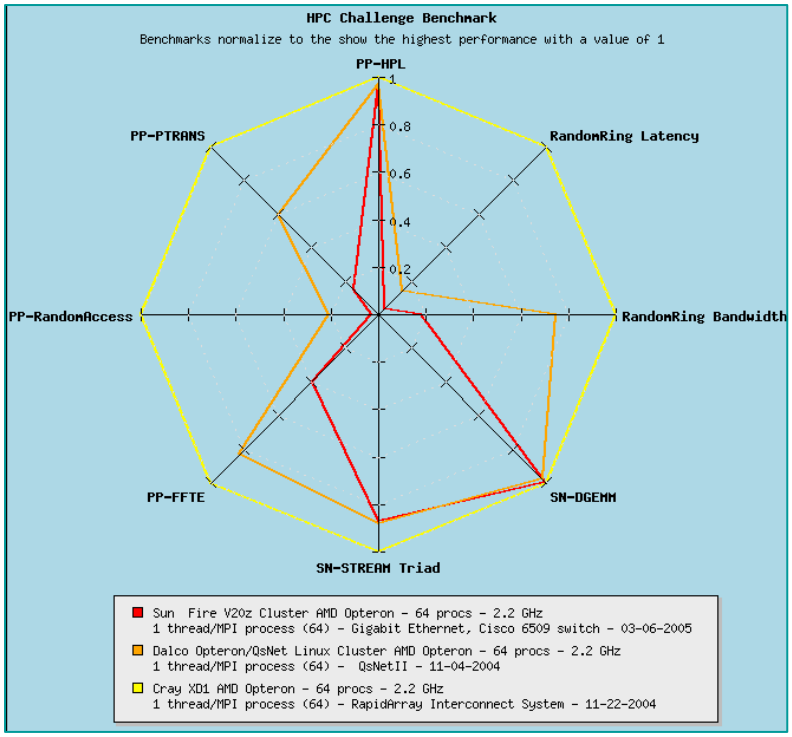
FLOPS alone are not enough.

Appropriate benchmarks needed to rank/bring visibility to more balanced machines critical for today's applications.

Cray XD1 -- Custom Interconnect

Dalco Linux Cluster -- Quadrics I nterconnect

Sun Fire Cluster -- Gigabit ethernet Interconnect



Three systems using the same processor and number of processors.

- *AMD Opteron 64 processors 2.2 GHz*
- ***Difference is in way the processors are interconnected***

HPC Challenge benchmarks measure different machine characteristics

- ***Linpack*** *and* ***matrix multiply*** *are computationally intensive*
- ***PTRANS*** *(matrix transpose),* ***RandomAccess*** **,** *bandwidth/latency tests and other tests begin to reflect stress on memory system*

Information courtesy of Jack Dongarra

An Integrated Resource Environment Needed to Support Data-Oriented Applications

SDSC HIGH PERFORMANCE COMPUTING SYSTEMS

- **DataStar**
  - 15.6 TFLOPS Power 4+ system
  - 7.125 TB total memory
  - Up to 4 GBps I/O to disk
  - 115 TB GPFS filesystem
- **Blue Gene Data**
  - First academic IBM Blue Gene system
  - 17.1 TF
  - 1.5 TB total memory
  - 3 racks, each with 2,048 PowerPC processors and 128 I/O nodes
- **TeraGrid Cluster**
  - 524 Itanium2 IA-64 processors
  - 2 TB total memory
  - Also 16 2-way data I/O nodes

http://www.sdsc.edu/user_services/

SDSC DATA COLLECTIONS, ARCHIVAL AND STORAGE SYSTEMS

- 2.4 PB Storage-area Network (SAN)
- 25 PB StorageTek/IBM tape library
- HPSS and SAM-QFS archival systems
- DB2, Oracle, MySQL
- Storage Resource Broker
- Supporting servers: IBM 32-way p690s, 72-CPU SunFire 15K, etc.

Support for community data collections and databases

Data management, mining, analysis, and preservation

http://datacentral.sdsc.edu/

SDSC SCIENCE and TECHNOLOGY STAFF, SOFTWARE, SERVICES

- Data-oriented Community SW, toolkits, portals, codes
- DataCentral national hosting repository
- Chronopolis services (w/ UCSDL)
- Data User Services
- Application/Community Collaborations
- Education and Training

http://www.sdsc.edu/

Data Services – What do Users Want?

How do I make sure that my data will be there when I want it?

How can I combine my data with my colleague's data?

How should I organize my data?

How should I display my data?

What are the trends and what is the noise in my data?

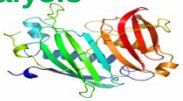
My data is confidential; how do I make sure that it is seen/used only by the right people?

How can I make my data accessible to my collaborators?

Services: Integrated Environment Is Key

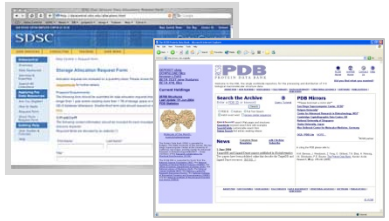
modeling

analysis



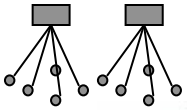
simulation

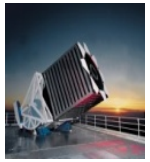
visualization



File systems,
Database systems,
Collection Management
Data Integration, etc.

instruments





Sensor-nets

computers

Many Data Sources

Data Access

Data Use

Data Management

Data Storage

Integrated Infrastructure

- *Database selection and schema design*
- *Portal creation and collection publication*
- *Data analysis*
- *Data mining*
- *Data hosting*
- *Preservation services*
- *Domain-specific tools*
  - *Biology Workbench*
  - *Montage (astronomy mosaicking)*
  - *Kepler (Workflow management)*
- *Data visualization*
- *Data anonymization, etc.*

SDSC DataCentral: National Data Hosting Facilities

- **Broad program to support research and community data collections and databases**
- **DataCentral services include:**
  - Public Data Collections and Database Hosting
  - Long-term storage and preservation (tape and disk)
  - Remote data management and access (SRB, portals)
  - Data Analysis, Visualization and Data Mining
  - Professional, qualified 24/7 support

Web-based portal access

- **DataCentral resources include**
  - 1 PB On-line disk
  - 25 PB StorageTek tape library capacity
  - 540 TB Storage-area Network (SAN)
  - DB2, Oracle, MySQL
  - Storage Resource Broker
  - Gpfs-WAN with 700 TB

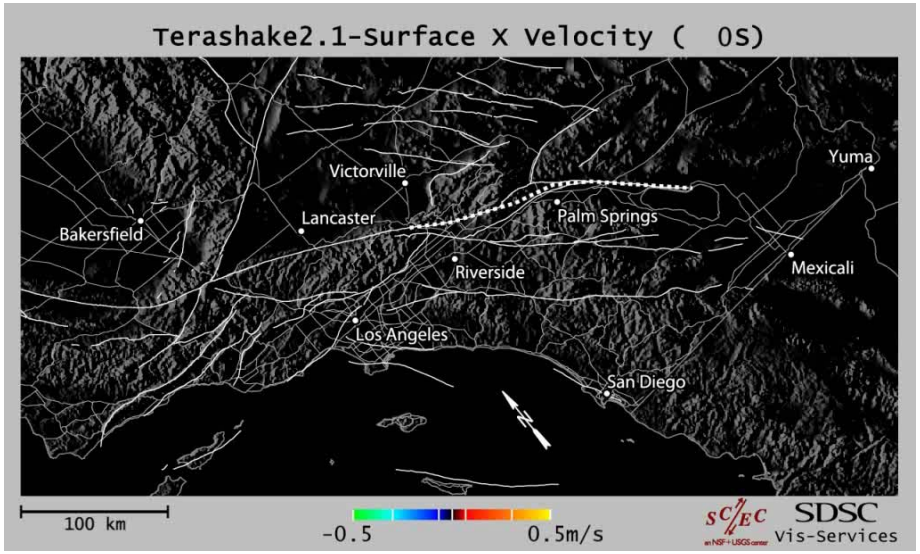
DataCentral Allocated Collections include

Seismology	3D Ground Motion Collection for the LA Basin
Atmospheric	Sciences50 year Downscaling of Global Analysis over California Region
Earth Sciences	NEXRAD Data in Hydrometerology and Hydrology
Elementary Particle Physics	AMANDA data
Biology	AfCS Molecule Pages
Biomedical Neuroscience	BIRN
Networking	Backbone Header Traces
Networking	Backscatter Data
Biology	Bee Behavior
Biology	Biocyc (SRI)
Art	C5 landscape Database
Geology	Chronos
Biology	CKAAPS
Biology	DigEmbryo
Earth Science Education	ERESE
Earth Sciences	UCI ESMF
Earth Sciences	EarthRef.org
Earth Sciences	ERDA
Earth Sciences	ERR
Biology	Encyclopedia of Life

Life Sciences	Protein Data Bank
Geosciences	GEON
Geosciences	GEON-LIDAR
Geochemistry	Kd
Biology	Gene Ontology
Geochemistry	GERM
Networking	HPWREN
Ecology	HyperLter
Networking	IMDC
Biology	Interpro Mirror
Biology	JCSG Data
Government	Library of Congress Data
Geophysics	Magnetics Information Consortium data
Education	UC Merced Japanese Art Collections
Geochemistry	NAVDAT
Earthquake Engineering	NEESIT data
Education	NSDL
Astronomy	NVO
Government	NARA
Anthropology	GAPP

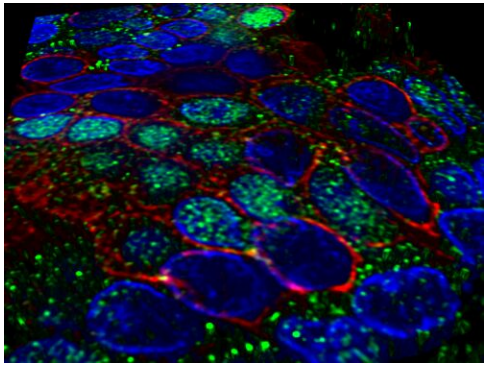
Neurobiology	Salk data
Seismology	SCEC TeraShake
Seismology	SCEC CyberShake
Oceanography	SIO Explorer
Networking	Skitter
Astronomy	Sloan Digital Sky Survey
Geology	Sensitive Species Map Server
Geology	SD and Tijuana Watershed data
Oceanography	Seamount Catalogue
Oceanography	Seamounts Online
Biodiversity	WhyWhere
Ocean Sciences	Southeastern Coastal Ocean Observing and Prediction Data
Structural Engineering	TeraBridge
Various	TeraGrid data collections
Biology	Transporter Classification Database
Biology	TreeBase
Art	Tsunami Data
Education	ArtStor
Biology	Yeast regulatory network
Biology	Apoptosis Database
Cosmology	LUSciD

Data Visualization

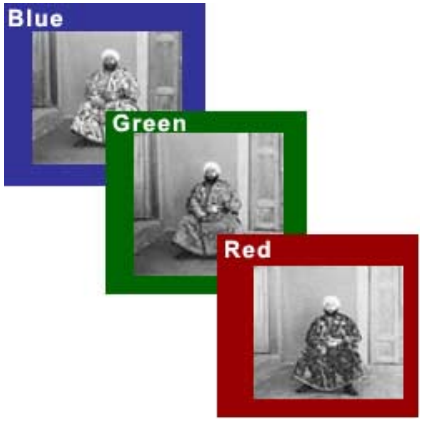


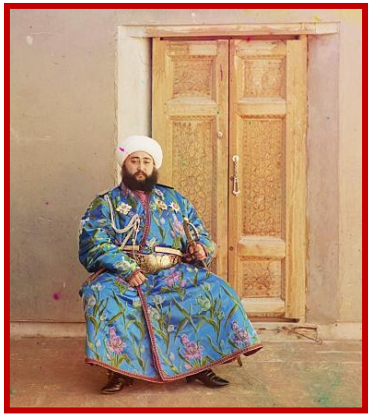
SCEC Earthquake simulations

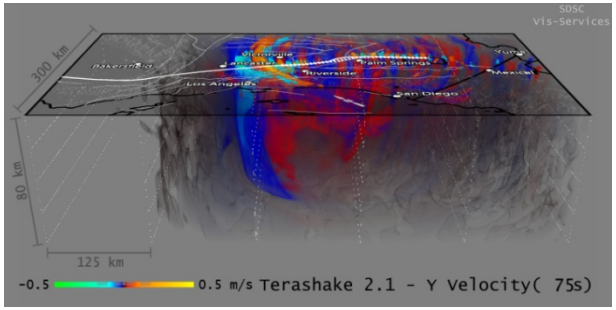
Visualization of Cancer Tumors



Prokudin– Gorskii historical images







Information and images courtesy of Amit Chourasia, SCEC, Steve Cutchin, Moores Cancer Center, David Minor, U.S. Library of Congress

Building Successful Cyberinfrastructure

The page cannot be found

The page you are looking for might have been removed, had its name changed, or is temporarily unavailable.

Please try the following:

- If you typed the page address in the Address bar, make sure that it is spelled correctly.
- Open the www2.hawaii.edu home page, and then look for links to the information you want.
- Click the Back button to try another link.
- Click Search to look for information on the Internet.

HTTP 404 - File not found
Internet Explorer

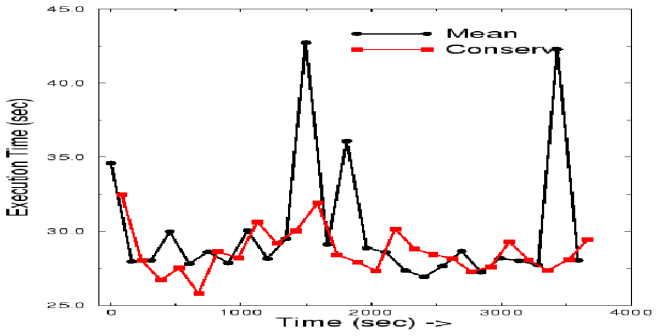
"Good" Data Cyberinfrastructure incorporates the "ilities"

- Scalability
- Interoperability
- Reliability
- Capability
- Sustainability
- Predictability
- Accessibility
- Responsibility
- Accountability
- …

Entity at risk	What can go wrong	Frequency
File	Corrupted media, disk failure	1 year
Tape	+ Simultaneous failure of 2 copies	5 years
System	+ Systemic errors in vendor SW, or malicious user, or operator error that deletes multiple copies	15 years
Archive	+ Natural disaster, obsolescence of standards	50 - 100 years

Data Reliability: What can go wrong

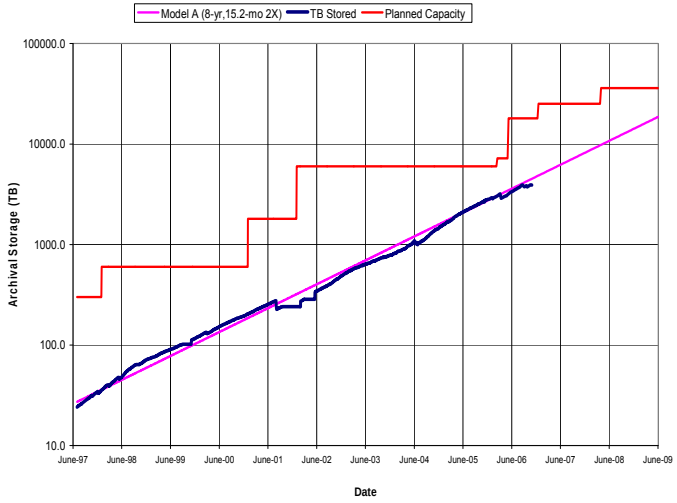
Predictable performance critical for user planning and optimization



Information courtesy of Reagan Moore, Jennifer Schopf

Good Data Infrastructure Incurs Real Costs

Capacity Costs



- *Most valuable data must be replicated*
- *SDSC research collections have been doubling every 15 months.*
- ***SDSC storage** is 25 PB and counting. Data is from supercomputer simulations, digital library collections, etc.*

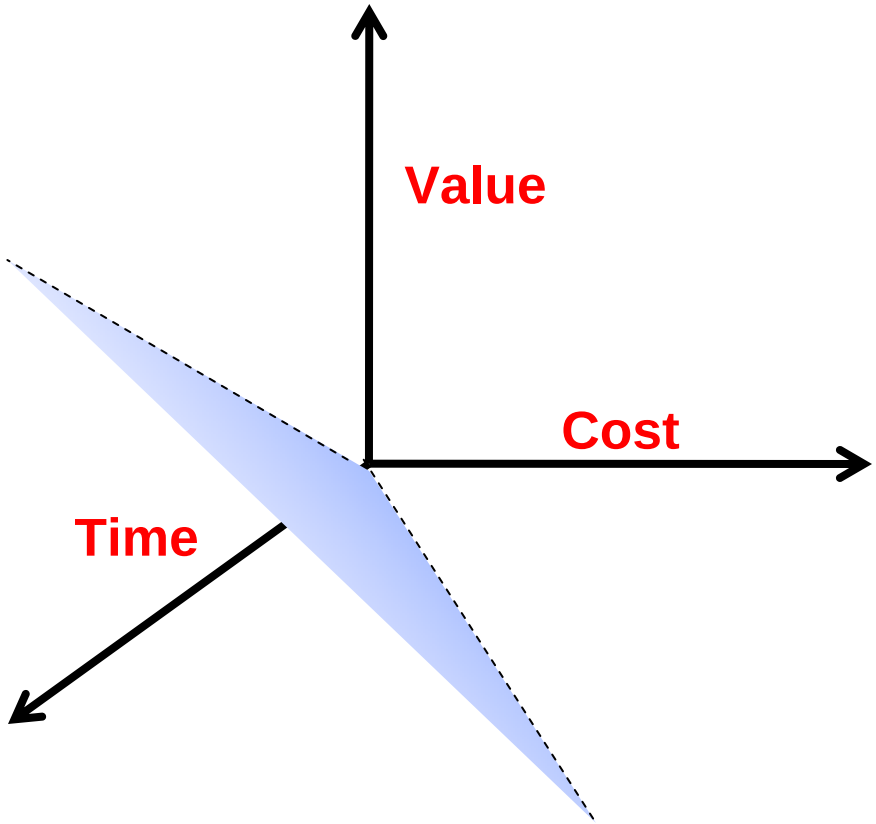
Capability Costs

- ***Reliability** increased by up-to-date and robust hardware and software for*
  - *Replication (disk, tape, geographically)*
  - *Backups, updates, syncing*
  - *Audit trails*
  - *Verification through checksums, physical media, network transfers, copies, etc.*
- ***Data professionals needed** to facilitate*
  - *Infrastructure maintenance*
  - *Long-term planning*
  - *Restoration, and recovery*
  - *Access, analysis, preservation, and other services*
  - *Reporting, documentation, etc.*

Information courtesy of Richard Moore

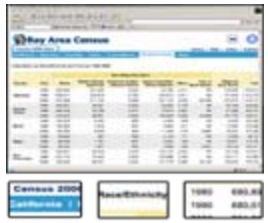
Economic Sustainability

- **Data Preservation** is the Grand Challenge for economically sustainable Cyberinfrastructure
- Good preservation requires **continuous support**
- **Key questions:**
  - What should we save?
  - Who should save it?
  - How should we save it?
  - Who should have access to it?
  - Who should pay for it?



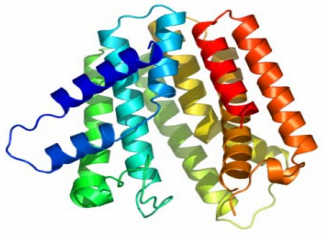
What Should We Save?

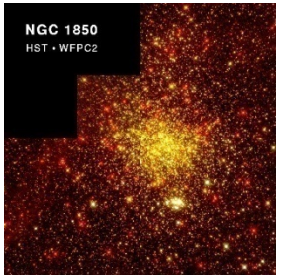
Data we* want to keep over the long-term:



Sarbanes-Oxley
Financial and Accounting Disclosure Information

- **We = "Society"**
  - Official and historically valuable data (Census information, presidential emails, Shoah Collection, etc.)
- **We = Research Community**
  - Protein Data Bank, National Virtual Observatory, etc.
- **We = Me**
  - My medical record, my Quicken data, digital photos of my Mom's 80th birthday, etc.





Who Pays? The "Free Rider" Non-Solution

- **Inadequate/unrealistic approach: "Let X do it" where X is:**
  - The Government
  - The Libraries
  - The Archivists
  - Google
  - Data users
  - Data owners
  - Data creators, etc.

- **Creative partnerships needed** to provide preservation solutions with
  - Trusted stewards
  - Feasible costs for users
  - Sustainable costs for infrastructure
  - Very low risk for data loss, etc.

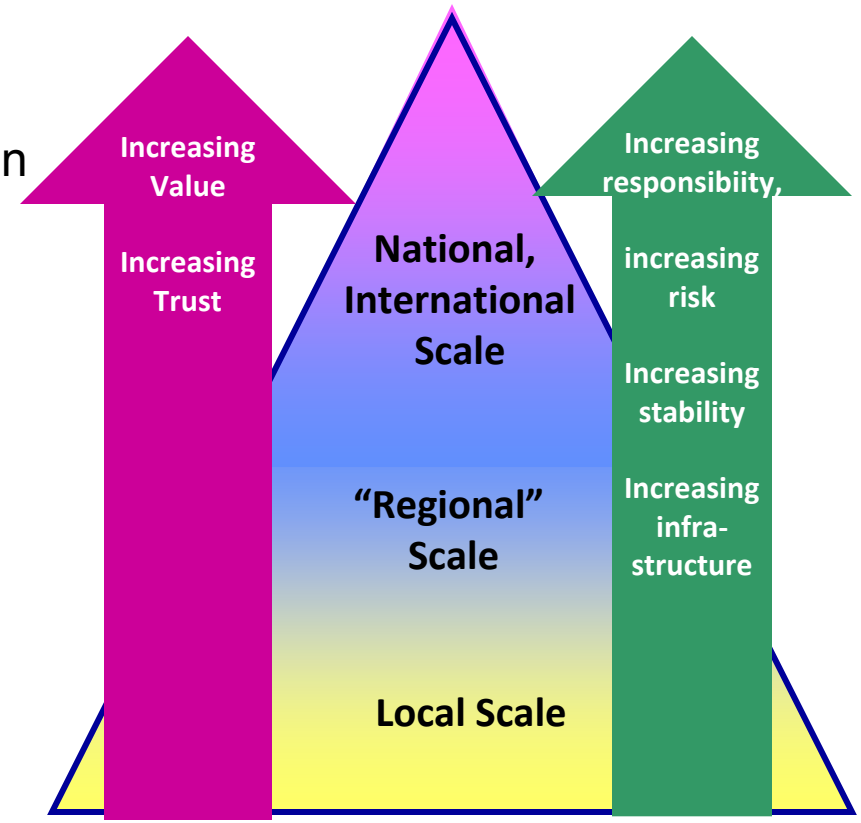
A Framework for Digital Stewardship and Preservation

Digital Data Collections

Reference, nationally/internationally important, irreplaceable data collections

Key research and community data collections

Personal data collections



Repositories / Facilities

National / international scale repositories, libraries, archives

"Regional" scale libraries and targeted data archives / centers

Private repositories

The Data Pyramid

Who Pays? Multiple Solutions

The LIBRARY of CONGRESS

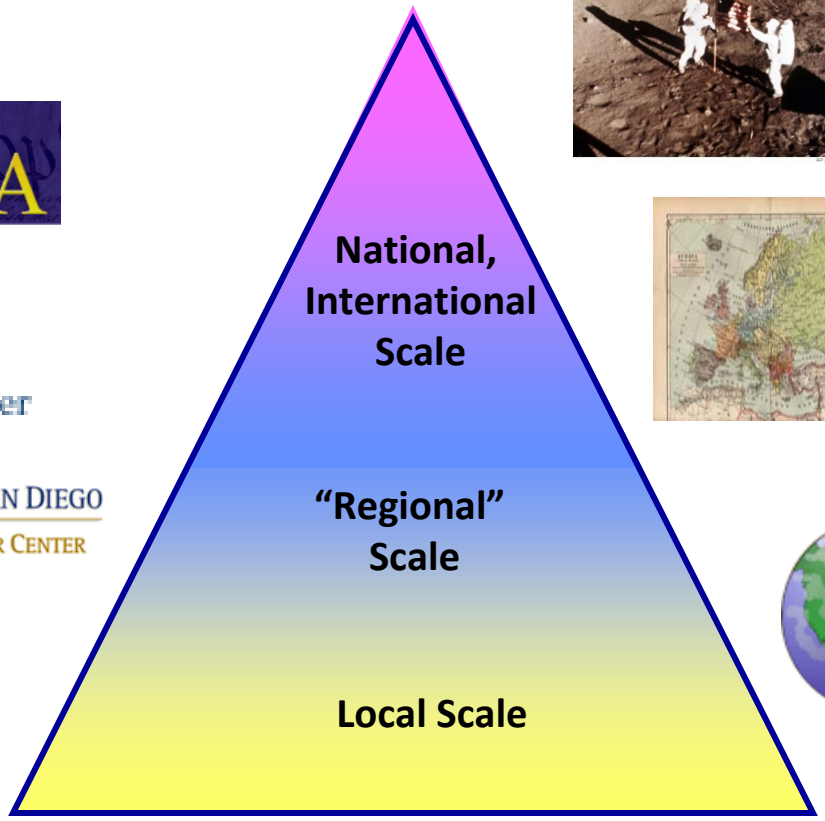
NARA

Georgia Tech Library & Information Center

UNIVERSITY of CALIFORNIA, SAN DIEGO MEDICAL CENTER MOORES CANCER CENTER

amazon web services™

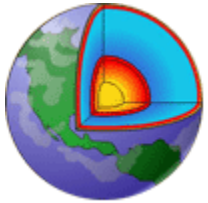
Google™

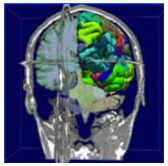


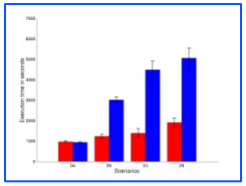
The Data Pyramid

UniProt the universal protein resource

PDB PROTEIN DATA BANK

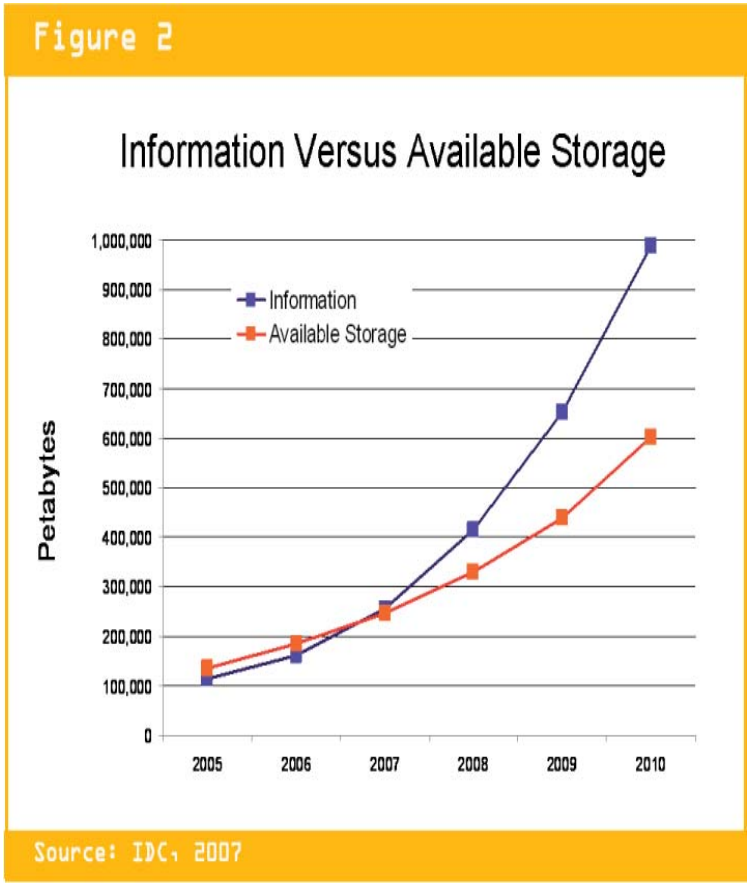






The Data Problem is About to Get Worse

- **2007 is the "crossover year"** where the amount of digital information is greater than the amount of available storage
- **Increasing requirements for data retention in private, public, academic sectors render data infrastructure and preservation policies more critical**
  - Agency requirements for data management and preservation, Sarbaanes-Oxley, etc.
- **Data Center systems / functionalities needed include**
  - Data Management & Movement
  - Metadata Catalog
  - Client Interfaces
  - Administration and policies
  - Authentication
  - Trust management, etc.



Sources: *"The Expanding Digital Universe: A forecast of Worldwide Information Growth through 2010" IDC Whitepaper, March 2007*

100 Years of Digital Data

- **Digital data is the natural resource of the Information Age**
  - Fragile– dependent on rapidly changing technologies
  - Valuable data lost often cannot be covered
- **100 Years represents**
  - Dozens of new generations of technologies
  - 100's of new data standards and formats in many communities
  - Thousands+ of new valued collections
  - Millions of potential users with as yet unknown information needs and workflows

What can we do?

- **Plan for preservation** throughout all stages of the digital life cycle – from creation to stewardship and beyond
- Ensure that **IT policies** include requirements that support preservation
- Ensure that **IT budgets** at all levels include appropriate costs for adequate levels of preservation
- Coordinate creative solutions throughout the data pyramid to **spread responsibility and cost**

Many Thanks

Chaitan Baru,Jennifer Schopf, Brian Lavoie, Brian Schottlaender, Mark Miller, Alex Szalay, Reagan Moore, Authors of the IDC Report, Ben Tolo, Richard Moore, Mike Norman, David Minor, Amit Chourasia, Jack Dongarra, Natasha Balac, U.S. Library of Congress, Moores Cancer Center, National Archives and Records Administration, NSF, Southern California Earthquake Center, Chris Greer, Steve Cutchin, UCAR, NVO, NASA, and many others

berman@sdsc.edu